DasSarma et al. Saline Systems 2010, 6:12
http://www.salinesystems.org/content/6/1/12

SALINE SYSTEMS

DATABASE Open Access

HaloWeb: the haloarchaeal genomes database

Satyajit L DasSarma[1], Melinda D Capes[1,2], Priya DasSarma[1], Shiladitya DasSarma[1,2*]

Abstract

Background: Complete genome sequencing together with post-genomic studies provide the opportunity for a comprehensive 'systems biology' understanding of model organisms. For maximum effectiveness, an integrated database containing genomic, transcriptomic, and proteomic data is necessary.

Description: To improve data access and facilitate functional genomic studies on haloarchaea in our laboratory, a dedicated database and website, named HaloWeb, was developed. It incorporates all finished and publicly released haloarchaeal genomes, including gene, protein and RNA sequences and annotation data, as well as other features such as insertion element sequences. The HaloWeb database was designed for easy data access and mining, and includes tools for tasks such as genome map generation, sequence extraction, and sequence editing. Popular resources at other sites, *e.g.*, NCBI PubMed and BLAST, COG and KOG protein clusters, KEGG pathways, and GTOP structures were dynamically linked. The HaloWeb site is located at http://halo4.umbi.umd.edu, and at a mirror site, http://halo5.umbi.umd.edu, with all public genomic data and NCBI, KEGG, and GTOP links available for use by the academic community. The database is curated and updated on a regular basis.

Conclusions: The HaloWeb site includes all completely sequenced haloarchaeal genomes from public databases. It is currently being used as a tool for comparative genomics, including analysis of gene and genome structure, organization, and function. The database and website are up-to-date resources for researchers worldwide.

Background

Genomic data are essential resources for modern biology and are most useful when freely accessible to all. This is especially true when databases are curated and simple and efficient data mining tools are available. Major centralized repositories have been useful, and play a crucial role [1-5]. However, due to the complexity and diversity of genomic data, it is very difficult, if not impossible, to meet all scientific demands solely through these major repositories. Well-designed smaller, more specific (clade or family) databases and websites can be vital for analysis and research, especially for individual laboratories focusing on model organisms [6].

The first haloarchaeal genome sequenced was that of *Halobacterium* sp. NRC-1 [7,8]. Initially, the 191 kilobase pair plasmid pNRC100 was sequenced and made public in 1998 [7]. In 2000, with the sequencing of the remainder of the 2.57 megabase pair genome of NRC-1, the annotation of pNRC100 was extensively revised and updated [8]. To provide access to the most current data and facilitate functional genomic studies on *Halobacterium* sp. NRC-1, we created a custom database and website named HaloWeb. The prototype HaloWeb site was made available to the public in 2000 as a service to the community and has been available for the past ten years [6-21].

With the recent increase in the number of completed genomes, including ten additional haloarchaeal genomes [22-28], research efforts have shifted from the single- to the multiple-genome level. As a result, it became necessary to update the HaloWeb site to incorporate the newly sequenced genomes, including up-to-date annotation data. The updated HaloWeb site incorporates enhanced data access and mining tools for *Halobacterium* sp. NRC-1 and the other haloarchaeal genomes.

Among the onsite tools are those for genome map generation, gene and intergenic sequence extraction, and sequence editing, which have been developed and implemented on the website. In addition, other popular web tools and resources have been dynamically linked. The database and website also provide templates for additional on-going genome sequencing projects, and

* Correspondence: sdassarma@som.umaryland.edu
[1]Department of Microbiology & Immunology, University of Maryland School of Medicine, Baltimore, MD, USA
Full list of author information is available at the end of the article

we expect to maintain and update resources for future data mining. Finally, the HaloWeb platform also provides an information management system to our laboratory for integration of public genomic data with additional proprietary transcriptomic and comparative genomic resources.

Results and Discussion

The HaloWeb server has been established utilizing Free/Libre and Open Source Software (FLOSS) including the Linux, Apache, MySQL, and Perl (LAMP) stack [29]. The HaloWeb gateway page (Figure 1) contains links to the 11 haloarchaeal homepages, as well as other useful resources such as HaloEd, a database for education using halophilic microorganisms, and convention and conversion information. Most information is freely accessible in the public domain portion of the site.

HaloWeb Genome Home

The genome homepage for each organism contains links to the organism's gene table, search page, and genome maps, along with the sequence editing tool, links to BLAST and genome sequence download pages in NCBI, as well as abstracts on the organism in PubMed.

Gene Table

The gene table allows for genomic analysis of all 11 organisms by providing selection options using different criteria, such as replicon and gene type. Having a uniform interface for interaction also generates a consistent view, from which database transversal is facilitated. The gene table contains data for locus, orientation, replicon, annotation, and gene ID, for each gene.

Search Tool

The search tool provides a comprehensive approach to data mining, allowing a search for genes based on ID number, name, annotation, or location in each genome. This is implemented using MySQL queries to the organism's database, optimized for quick retrieval by using the minimum columns necessary to complete the table, in a unified interface.

Gene Page

The HaloWeb gene pages (Figure 2) allow access, via links, to information resources for the gene using our custom query interface tools to the database. The tools permit BLASTing the gene against protein and nucleotide databases at NCBI, accessing protein data at GenBank [2], and accessing the associated COGs and KOGs from NCBI [30]. There are also links to the KEGG [3] and GTOP [4] databases. A table is also generated containing links to surrounding genes, the number of which may be selected by a dropdown menu. The table also contains each gene's ID, name, size, and annotation.

For an alternate way of navigating the database, a gene map with links to surrounding genes is available. The number of genes is regulated by the dropdown menu, and uses an image map to add informational popups and links to the otherwise static map. Controls below the map move the gene map window by changing the gene selected or by allowing leaps to either end of the current map. Below the map is a form containing controls for a popup with sequence data for the current gene region. Optionally, sequence data for an area around the gene, including intergenic sequences, can be retrieved.

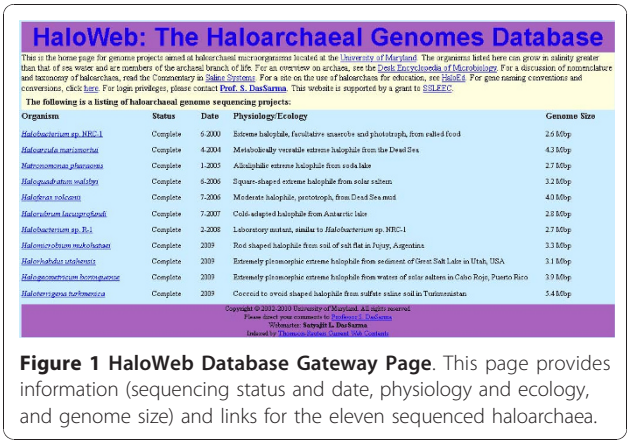

Figure 1 HaloWeb Database Gateway Page. This page provides information (sequencing status and date, physiology and ecology, and genome size) and links for the eleven sequenced haloarchaea.

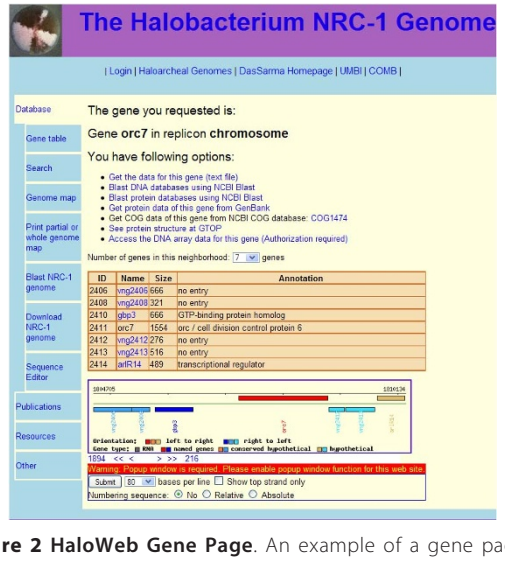

Figure 2 HaloWeb Gene Page. An example of a gene page is shown for *Halobacterium* sp. NRC-1 *orc7*. A variety of information (when available) is linked near the top of the page, followed by a table of the gene region, a corresponding genetic map, and sequence data form.

Maps

Map queries are also possible in HaloWeb (Figure 3). The first dialog is accessed by clicking on the "Genome Map" link. This dialog contains a replicon selection radio box and a button to continue to the next section. The second section is a form to set the format for the generation of the map, including dropdown menus for bases per line, pixels per line, and a list of genes. The list of genes is used for selecting the first and last genes, using buttons to fill in the read-only text boxes. There are also check boxes to use links or get the entire replicon. Finally, the map is generated by clicking the "Submit" button, which opens a new tab with the image.

Conclusions

With the completion of the updated HaloWeb site, genome data from a major family of microorganisms, the haloarchaea, are readily accessible. This resource has served the academic research community for many years. In addition, HaloWeb also includes proprietary in-house generated data, including microarray and protein cluster data, and serves as a useful laboratory information management system [31].

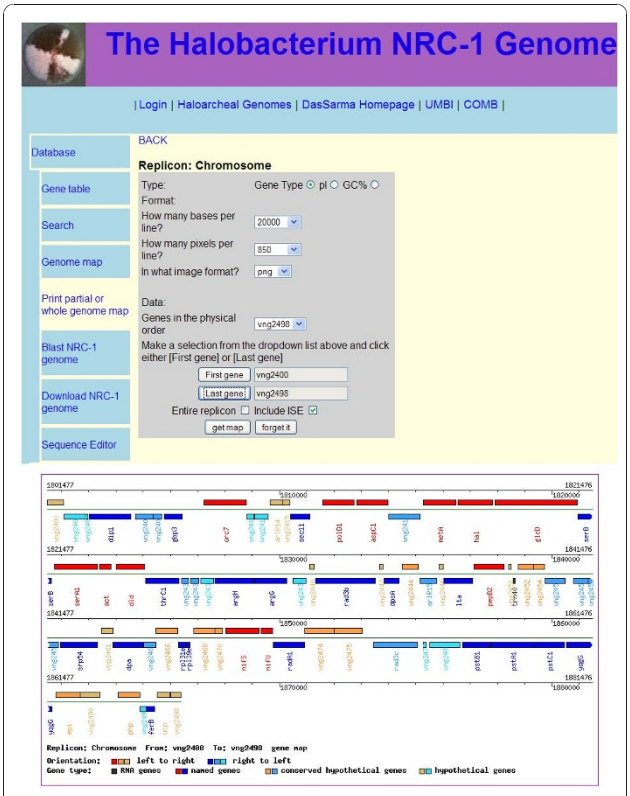

Figure 3 HaloWeb Map Query Page. An example of a map query is shown for the *Halobacterium* sp. NRC-1 chromosome. After input of the desired specifications, a map of the region is generated.

Methods

Software Tools

Red Hat Enterprise and Fedora Linux, in both 32 bit and 64 bit versions, are used to run the servers. The Apache 2 web server is used to serve up web pages, and a MySQL Community server is used for the database backend. Most scripts are implemented using Perl, connecting to MySQL using the DataBase Independence (DBI) Perl module from Common Perl Archive Network (CPAN) as our database frontend, to allow the greatest flexibility in script writing and database program usage. The usage of the Perl language allows easy graphics generation by the GD library, such as the gene mapping utility, through the GD object-oriented module, and parameter passing is through the Common Gateway Interface (CGI) module. In some cases, JavaScript code is also utilized.

Genome data

Genome data for the following organisms was obtained from NCBI: *Halobacterium* sp. NRC-1, *Haloarcula marismortui*, *Natronomonas pharaonis*, *Haloquadratum walsbyi*, *Haloferax volcanii*, *Halorubrum lacusprofundi*, *Halobacterium* sp. R-1, *Halomicrobium mukohataei*, *Halorhabdus utahensis*, *Halogeometricum borinquense*, and *Haloterrigena turkmenica*.

Acknowledgements

We thank Professor Yanhe Ma for organizing the 9th International Conference on Halophilic Microorganisms in Beijing, China, where this work was originally presented [19]. We also thank Peijun Zhang, Jeetendra Soneja, and Jeff Kumar for their work to establish the HaloWeb prototype. This work was supported by the Henry M. Jackson Foundation grant HU0001-09-1-0002-660883, the National Aeronautics and Space Administration grant NNX09AC68G, and the National Science Foundation grant MCB-0450695 to S.D.

Author details

[1]Department of Microbiology & Immunology, University of Maryland School of Medicine, Baltimore, MD, USA. [2]Graduate Program in Life Sciences, University of Maryland School of Medicine, Baltimore, MD, USA.

Authors' contributions

The HaloWeb database structure was designed by SD, and software development and systems administration was provided by SLD. Database tables and website testing were provided by MDC and PD. The manuscript was written by SLD, MDC, PD, and SD. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 3 November 2010 Accepted: 30 December 2010
Published: 30 December 2010

References

1. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2010, **38**:5-16.

2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2010, **38**:46-51.
3. Tatusova T: **Genomic databases and resources at the National Center for Biotechnology Information.** *Methods Mol Biol* 2010, **604**:17-44[http://www.springerlink.com/content/p1378813006754uh/#section=671303&page=1].
4. Fukuchi S, Homma K, Sakamoto S, Sugawara H, Tateno Y, Gojobori T, Nishikawa K: **The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions.** *Nucleic Acids Res* 2009, **3**:333-337.
5. Aoki KF, Kanehisa M: **Using the KEGG database resource.** *Curr Protoc Bioinformatics* 2005, **Chapter 1**:Unit 1.12.
6. DasSarma S: **Genome sequence of an extremely halophilic archaeon.** In *Microbial Genomes.* Edited by: Read T, Nelson KE, Fraser CM. Totowa: Humana Press, Inc; 2004:383-399.
7. Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J, Hood L, DasSarma S: **Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome?** *Genome Res* 1998, **8**:1131-1141.
8. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, *et al*: **Genome sequence of *Halobacterium* species NRC-1.** *Proc Natl Acad Sci USA* 2000, **97**:12176-12181.
9. Boucher Y, Huber H, L'Haridon S, Stetter KO, Doolittle WF: **Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders *Thermoplasmatales* and *Archaeoglobales*.** *Mol Biol Evol* 2001, **18**:1378-1388.
10. Liang P, Labedan B, Riley M: **Physiological genomics of *E. coli* protein families.** *Physiol Genomics* 2002, **9**:15-26.
11. Levin I, Giladi M, Altman-Price N, Ortenberg R, Mevarech M: **An alternative pathway for reduced folate biosynthesis in bacteria and halophilic archaea.** *Mol Microbiol* 2004, **54**:1307-1318.
12. Lichi T, Ring G, Eichler J: **Membrane binding of SRP pathway components in the halophilic archaea *Haloferax volcanii*.** *Eur J Biochem* 2004, **271**:1382-1390.
13. Mizuki T, Kamekura M, DasSarma S, Fukushima T, Usami R, Yoshida Y, Horikoshi K: **Ureases of extreme halophiles of the genus *Haloarcula* with a unique structure of gene cluster.** *Biosci Biotechnol Biochem* 2004, **68**:397-406.
14. Nichols DS, Miller MR, Davies NW, Goodchild A, Raftery M, Cavicchioli R: **Cold adaptation in the Antarctic archaeon *Methanococcoides burtonii* involves membrane lipid unsaturation.** *J Bacteriol* 2004, **186**:8508-8515.
15. Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P: **Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to *Thermococcales*?** *Genome Biol* 2005, **6**:42.
16. Oren A, Larimer F, Richardson P, Lapidus A, Csonka LN: **How to be moderately halophilic with broad salt tolerance: clues from the genome of *Chromohalobacter salexigens*.** *Extremophiles* 2005, **9**:275-279.
17. Lledó B, Marhuenda-Egea FC, Martínez-Espinosa RM, Bonete MJ: **Identification and transcriptional analysis of nitrate assimilation genes in the halophilic archaeon *Haloferax mediterranei*.** *Gene* 2005, **361**:80-88.
18. Bab-Dinitz E, Shmuely H, Maupin-Furlow J, Eichler J, Shaanan B: ***Haloferax volcanii* PitA: an example of functional interaction between the Pfam chlorite dismutase and antibiotic biosynthesis monooxygenase families?** *Bioinformatics* 2006, **22**:671-675.
19. Ma Y, Galinski EA, Grant WD, Oren A, Ventosa A: **Halophiles 2010: Life in Saline Environment.** *Appl Environ Microbiol* 2010, **76**:6971-6981.
20. Perez-Rueda E, Janga SC: **Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin.** *Mol Biol Evol* 2010, **27**:1449-1459.
21. Zafrilla B, Martínez-Espinosa RM, Esclapez J, Pérez-Pomares F, Bonete MJ: **SufS protein from *Haloferax volcanii* involved in Fe-S cluster assembly in haloarchaea.** *Biochim Biophys Acta* 2010, **1804**:1476-1482.
22. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, Shannon P, Chiu Y, Weng RS, Gan RR, *et al*: **Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea.** *Genome Res Vol* 2004, **14**:2221-2234.
23. Hartman AL, Norais C, Badger JH, Delmas S, Haldenby S, Madupu R, Robinson J, Khouri H, Ren Q, Lowe TM, *et al*: **The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon.** *PLoS One* 2010, **5**:9605.
24. Malfatti S, Tindall B, Schneider S, Fähnrich R, Lapidus A, Labutti K, Copeland A, Glavina del Rio T, Nolan M, Chen F, *et al*: **Complete genome sequence of *Halogeometricum borinquense* type strain (PR3^T).** *Standards in Genomic Sciences* 2009, **1.**
25. Tindall BJ, Schneider S, Lapidus A, Copeland A, Rio TGD, Nolan M, Lucas S, Chen F, Tice H, Cheng JF, *et al*: **Complete genome sequence of *Halomicrobium mukohataei* type strain (arg-2^T).** *Standards in Genomic Sciences* 2009, **1.**
26. Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, *et al*: **Evaluation of three automated genome annotations for *Halorhabdus utahensis*.** *PLoS One* 2009, **4**:6291.
27. **NCBI.** [http://www.ncbi.nlm.nih.gov/].
28. Falb M, Pfeiffer F, Palm P, Rodewald K, Hickmann V, Tittor J, Oesterhelt D: **Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*.** *Genome Res* 2005, **15**:1336-1343.
29. Rosebrock E, Filson E: *Setting up LAMP* London: Sybex; 2004.
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
31. Capes MD, Coker JA, Gessler R, Grinblat-Huse V, DasSarma SL, Jacob CG, Kim JM, DasSarma P, DasSarma S: **The Information Transfer System of Halophilic Archaea.** *Plasmid* 2010.

doi:10.1186/1746-1448-6-12
Cite this article as: DasSarma *et al.*: HaloWeb: the haloarchaeal genomes database. *Saline Systems* 2010 **6**:12.